

本周在论文上的进展不够顺利。先前试图在 Hadoop 集群上实现社交网络的可视化受阻。

最近一直参加张老师的讨论班，从中了解了 hadoop 的基本原理以及编程和管理的细节。我们参考的书籍是 2009 年 Tom White 编写的《Hadoop 权威指南（第二版）》其中包含了一些成功应用 Hadoop 的实际案例。然而，就在周五在看最新版（第三版）时，在应用案例中发现 infochimps 的 Philip Kromer 利用 hadoop 对社交网络分析并可视化的案例，Using Pig and Wukong to Explore Billion-edge Network Graphs。

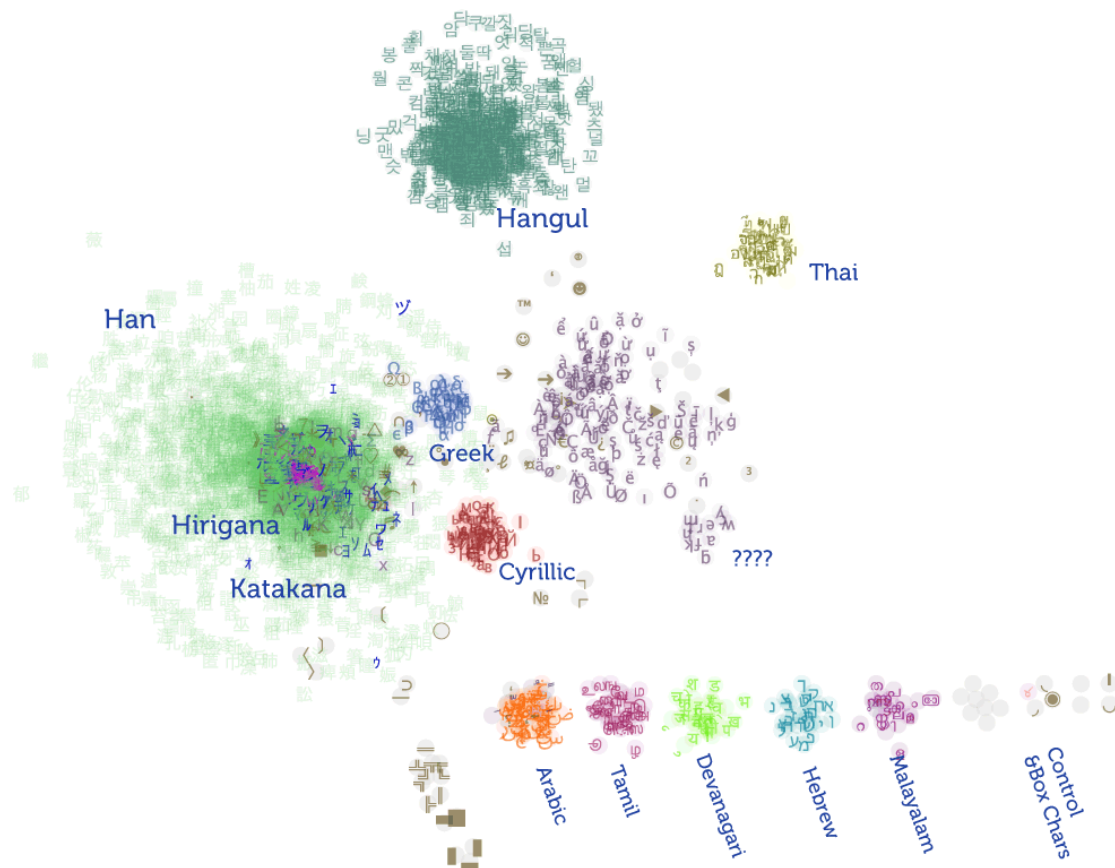


Figure 1 Twitter language map

图 1 表示的是 Twitter 中用户常用词的搭配情况，例如汉字中“最”和“近”是常常搭配在一起的。可视分析的过程：确定节点→节点之间的关系→聚类→可视化。

图 2 是对 2 亿条 twitter 消息分析之后绘制结果，具体过程如下：

1、分析社团：

首先通过“@”字符确定用户之间的回复情况，通过回复确定每个用户
通过 pig 查找用户之间的关联是否对称
计算每个节点的度，确定邻居节点，提取社团

2、可视化

根据每个节点的关系生成三角形，根据三角形计算聚类因子（不明白具体过程），最后对聚类的结果进行绘制，

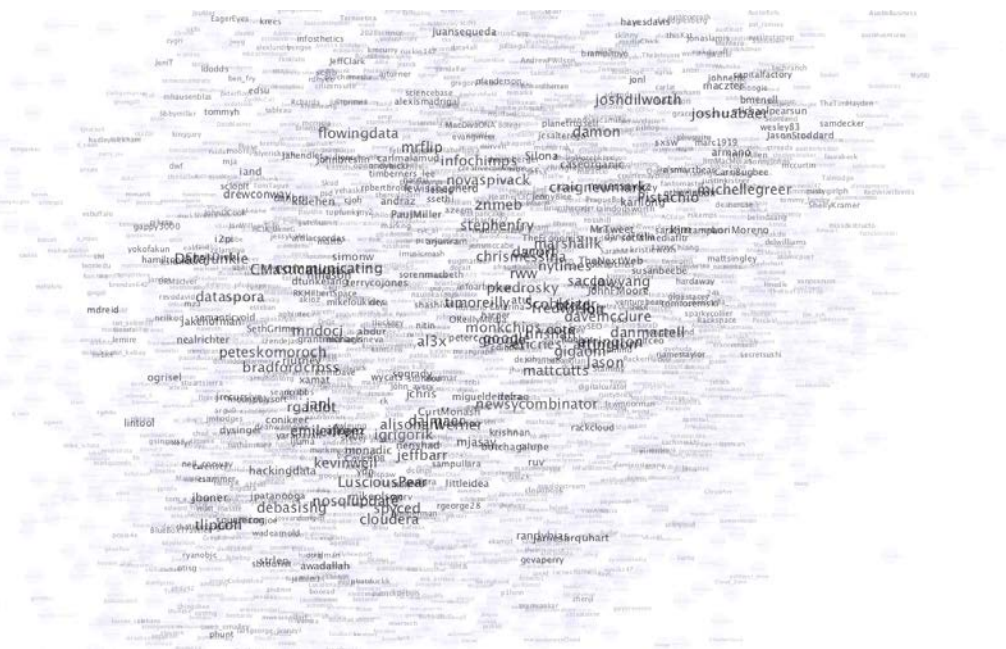


Figure 2 Big data community on Twitter

看到这个案例，只能说再次证明了大数据与可视化确实是当前研究的热点。当然，更多的是又一次失落，本来这是一个非常好的想法，但是现在又被别人抢先了。当然这里仍然有可以进一步探索的余地，就是如何提高分析和可视化的效率。案例最后提到现实数据中，数据量非常庞大，例如当红明星@britneyspears (5.2 million followers, 420,000 following as of July 2010) or @WholeFoods (1.7 million followers, 600,000 following)。同时大图又是非常稀疏的，几乎所有的数据都可以丢弃。

利用 **hadoop** 分析和计算，最重要的目标是如何让每个工作节点的处理时间比较均匀，也就是要对数据均匀划分。部分用户数据量可能非常大，部分用户的数据可以忽略不计，如何解决这个问题，是值得研究的。但是这个已经超出可视化的范畴，可能需要研究社交网络分析或者更多内容。或者说**研究这个问题是否有明确的意义，是否值得花费大量时间去研究？**

暂时没有明确的答案。周五我已着手另外的一个想法，就是对 **Twitter** 数据中的话题传播路径和话题演化过程进行可视化，算是对前面工作的进一步拓展，希望能得到比较好的可视化结果，能够赶上今年的 **Pacific VIS**（截至日期 9 月 2 日）。